STAT438 - ADVANCED RESEARCH METHODS IN BIOSTATISTICS

IMPLICATIONS OF MISSING DATA IN MULTIPLE REGRESSION ANALYSIS

April 11, 2019

Marcus Di Renzo, Student ID: 20552745
University of Waterloo
Department of Statistics & Actuarial Science

STAT438 Final Project

Contents

1 Introduction

1.1 Background

Within data collection and statistics, missing data is a pervasive issue that affects virtually all studies. Missing data can arise as a consequence of multiple sources, including item non-response within surveys, attrition in longitudinal cohort studies, missingness as a result of data storage issues, or as a result of the study design itself (Grace-Martin, 2018). Despite the ubiquity of missing data, it presents statistical challenges that are non-trivial to deal with, especially in circumstances where the extent of missing data is relatively large or affects multiple variables simultaneously. If missing data is ignored entirely, the resultant estimates we compute from our data sets may be biased, and therefore may not accurately represent the underlying quantity of interest. Other issues of ignoring missing data include a loss of statistical power, the extent to which the results represent the overall population, the computational complexity associated with the analysis, and the statistical inferences resulting from the study (Kang, 2013). In order to bypass this issue, analysts must leverage some form of statistical method, such as multiple imputation, single imputation, inverse probability weighting, or other methods in order to compute the missing observations themselves. This is a difficult problem, as we can never verify our results analytically from the observed data alone, and there is no guarantee the missing data itself is at all similar to the observed data. For example, in the case of categorical missing data, the missing data may be part of some unobserved category in our observed data set, which can create complications.

In addition to dealing with the missing data itself, the underlying reason for the missingness must also be considered. Formally, the missing data mechanism relates the observed covariates with the probability of missingness (Nakagawa, 2015). This is further complicated in scenarios in which multiple covariates contain missing values, as the underlying mechanism of missingness may vary between covariates and individuals. There are three major categories of missing mechanisms, including:

1) **Missing Completely at Random (MCAR)**: In this mechanism of missingness, the missing data is assumed to be missing entirely due to random chance, meaning there is no underlying mechanism of missingness. This means that the missingness is independent of the data we observe and the data we do not observe. This is the most trivial case to deal with, as if we analyze our complete data set (and ignore the missing data), we are still able to get unbiased estimations for our parameters of interest. In order to test whether the MCAR assumption is suitable, statistical tests such as *Little's MCAR Test* are often employed (Garson, 2015).

2) **Missing at Random (MAR)**: If the data is missing at random (also called missing conditionally at random), the missingness may depend on our observed data, but will not depend on the missing data itself. Cases such as this are more complex than the completely random case, but we can still employ many statistical methods such as imputation or inverse probability weighting to deal with these scenarios. If the data is truly MAR, then we cannot simply ignore the missingness, as this would lead to biased estimation results (Garson, 2015). Consequently, the missing data are typically modelled as some combination of observed covariates in order to identify suitable estimates for the missing values.

3) **Missing not at random (MNAR)**: If the data is missing non-randomly, this is the most complex case to deal with. In this scenario, the missing data depends on the missingness itself, even after considering the observed data we have. This means we cannot necessarily model the missing data as a function of the observed data, as these predictors may not be suitable and we lack information on covariates which may influence the missingness (Garson, 2015).

Therefore before deciding what method to apply to accommodate missing data, the underlying mechanism of missingness must also be examined to determine which methods are suitable in the context of the study. Without further research about the study design or subject-matter expertise related to the topic, this can be difficult to accomplish from the observed data alone.

1.2 Research Problem

The purpose of this report is to broadly examine the consequences of missing data on multiple regression. Though many different missing data strategies exist, there is no 'perfect' method of dealing with missing data due to the nature of the problem itself. Though much of previous literature often compares and contrasts overall classes of missing data methods, less attention is paid to the fine tuning of parameters within these classes themselves.

For example, in multiple imputation the analyst must decide the value of m to select, corresponding to the number of imputed datasets to be created. There is a trade-off in selecting this parameter - larger values of m should reduce the variability in our estimations, but can be more computationally complex, especially as the number of missing observations grows. Historic research in missing data suggested a lack of significant improvement in efficiency for selecting a value of m above 10 (Schafer, 1999), though many other researchers disputed this claim in future research. A 2007 study suggested 10-30% missing information should rely on 20 imputations (Graham et al., 2007) in order to produce more efficient standard errors, while other research suggests the number of imputations should be similar to the percentage of missing data (Royston et. al, 2011). From this we can see relative discrepancy among researchers, and there is no gold standard of imputed data sets to rely on (Allison, 2012).

Similarly, in single imputation (a simpler case of multiple imputation in which m is fixed to be 1) there is flexibility in the choice of model we use to impute the missing values based on our observed sample. Though previous research suggests non-parametric models are less suitable than parametric due to the larger bias in the imputed values (Binding, 2017), parametric models rely on the underlying assumption of correct model specification. This can lead to issues, as if the model is not correctly specified, inferences may be inaccurate or misleading. Non-parametric and semi-parametric modelling relax this assumption, and are thought to be more suitable and data-driven in the context of missing data (Silverman, 1985). These however, can be more complex to implement, especially for non-statisticians. Many non-parametric models require more investigation into suitable values of tuning parameters to best fit the data, which can be impractical or difficult under some circumstances.

Inverse probability weighting (henceforth called IPW) is another popular approach in dealing with missing data. The idea behind inverse probability weighting is to weight each subject by the inverse of the probability of being observed. This means that when we compute a regression model with these weights, the individuals who have a low probability of being completely observed are up-weighted in order to ensure the sample is more representative. IPW offers the advantage of being easy to program and relatively intuitive for individuals with non-statistics backgrounds, however this method relies on a restrictive set of assumptions which can be hard to verify in practice (Seaman & White, 2015). Notably, this technique relies on the assumption that the model used to estimate the weights is correctly specified, since otherwise the resultant estimates will be prone to bias in finite samples. As a result, much of recent literature focuses on using non-parametric models to estimate these weights, such as generalized boosting models (GBM) or random forest, though the majority of this research is focused on scenarios in which X (the independent covariates) is high dimensional. Similar to the above case, these models can also be difficult to fit or interpret for non-technical audiences.

This report will focus primarily on imputation and IPW in general, specified into distinct sub-categories. Single imputation will use the sample mean to impute missing values, and the drawbacks of this approach will be illustrated. These results will then be compared against IPW (with weights calculated using various approaches) and complete data analysis methods in order to highlight the consequences of using different missing data techniques. In the case of multiple imputation, the choice of m will be examined more thoroughly to determine the implications of selecting a poor choice of this value.

The primary focus of this report to determine the influence of different approaches to missing data on multiple regression modelling, particularly in terms of the final model fit, variable selection, and efficiency of the regression coefficients. The sample data used to highlight this will come from the Ozone data set in R. Much of missing data research is aimed at statistics specialists, and so missing data techniques are often not well understood by non-experts. By focusing specifically on multiple regression, model selection, and model fit, the focus is on a problem that is relevant to non-technical audiences. The result will be a more non-technical examination of missing data, the underlying mechanisms, and an intuitive way to apply each of these methods in a variety of contexts.

Though a similar problem was discussed in class, this research question is more complex as a result of the extent of missing data within the study data set. The Ozone data set, discussed in the following section, has missing data across multiple covariates. This is more reflective of a typical missing data scenario, whereas in class the focus was on a dataset with missingness in a single covariate. Additionally, the analysis conducted in class did not compare the model selection or overall model fit resulting from each method, which is a significant component of this report. Lastly, this report also further investigates the value of m used in multiple imputation; previous research has provided significant discrepancy in defining guidelines about how the value of m should be selected, and so this report investigates the results of previous research in context of real data, rather than simulations.

2 Data & Methodologies

2.1 Dataset

The dataset used in this report is the Ozone data set taken from the mlbench package in R. This data set has a total of 366 observations across a total of 13 covariates, some of which are not included analysis. This dataset corresponds to air pollution data collected within Los Angeles basin during the year of 1976. This data was collected over the course of 330 days, and the goal was to determine the relationship between ozone concentration (measured as the maximum one hour average each day) with other various meteorological quantities of interest (Breiman and Friedman, 1977). Note that ozone is used as the dependent variable in regression modelling, while the other covariates correspond to the independent covariates.

Each of the covariates used in the analysis are explained below. This information was predominantly collected from the package documentation and online reference tools. The names in brackets refer to their coding in the dataset, and a full visualization of each variable is available in Appendix B following the body of the report.

1) **Ozone**: Refers to the maximum ozone reading on the given measurement day, averaged over an hour long period and measured in parts-per-million (ppm). Ozone is a gas whose concentration is often of interest as it can have detrimental effects to human health, including inducing respiratory effects. Typically ozone levels are highest in the summer months, where sunlight is more direct and temperatures are hottest. Other factors such as wind, humidity, wind speed, and air pressure can influence ozone levels (Environmental Protection Agency, 2018).

2) **Pressure (Van_Pres)**: Measures the geopotential height at Vandenberg Air Force Base. This is a measurement in meters which corresponds to the height at which the atmospheric pressure is 500 millibars (a particular unit of pressure).

3) **Wind Speed (Wind_Sp)**: Measures the wind speed in miles per hour (mph), conducted at the Los Angeles International Airport (LAX).

4) **Humidity (Humid)**: Measures the humidity at LAX as a percentage value.

5) **Temperature_Sandburg (Temp_Sandburg)**: Measurement of the temperature at Sandburg, California measured in Fahrenheit (F).

6) **Temperature_El Monte (Temp_ElMonte)**: Measurement of the temperature at El Monte, California expressed in Fahrenheit (F).

7) **Height (Inv_Height)**: Measures the inversion base height at LAX, measured in feet. This is a particular height in which the temperature begins to increase with height, rather than decrease (WeatherStreet, 2010).

8) **Pressure Gradient (Pressure_Grad)**: This measurement was a gradient over the region from the Los Angeles International Airport (LAX) to Daggett, California. The resultant measurement is in millimeters of mercury (mmHg).

9) **Inversion Base Temperature (Inv_Temp)**: Measurements the inversion base temperature at the Los Angeles International Airport in Fahrenheit.

10) **Visibility**: The visibility measured in miles. In general, high values of ozone reduce the visibility (Environmental Protection Agency, 2018).

11) **Season**: Represents the season in which the measurement was recorded. Note that this variable was originally coded as month within the dataset, and has been redefined with months aggregated according to the corresponding season. As a consequence of the change in temperature and atmospheric conditions during seasons, ozone often displays seasonal variability (Environmental Protection Agency, 2018).

The following tables provide some diagnostic summaries of each of the covariates used in the analysis:

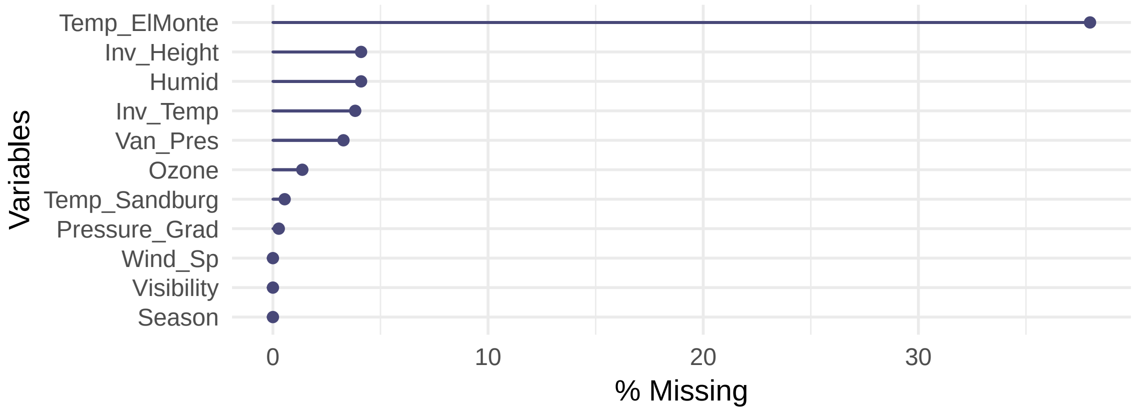
Table 1: Summary of Ozone Data Set (Variables 1-5)

Ozone	Van_Pres	Wind_Sp	Humid	Temp_Sandburg
Min. : 1.00	Min. :5320	Min. : 0.000	Min. :19.00	Min. :25.00
1st Qu.: 5.00	1st Qu.:5700	1st Qu.: 3.000	1st Qu.:49.00	1st Qu.:51.00
Median : 9.00	Median :5770	Median : 5.000	Median :65.00	Median :62.00
Mean :11.53	Mean :5753	Mean : 4.869	Mean :58.48	Mean :61.91
3rd Qu.:16.00	3rd Qu.:5830	3rd Qu.: 6.000	3rd Qu.:73.00	3rd Qu.:72.00
Max. :38.00	Max. :5950	Max. :11.000	Max. :93.00	Max. :93.00
NA's :5	NA's :12	NA	NA's :15	NA's :2

Table 2: Summary of Ozone Data Set (Variables 6-10)

Temp_ElMonte	Inv_Height	Pressure_Grad	Inv_Temp	Visibility
Min. :27.68	Min. : 111	Min. :-69.0	Min. :27.50	Min. : 0.0
1st Qu.:49.73	1st Qu.: 890	1st Qu.:-10.0	1st Qu.:51.26	1st Qu.: 70.0
Median :57.02	Median :2125	Median : 24.0	Median :62.24	Median :110.0
Mean :56.85	Mean :2591	Mean : 17.8	Mean :60.93	Mean :123.3
3rd Qu.:66.11	3rd Qu.:5000	3rd Qu.: 45.0	3rd Qu.:70.52	3rd Qu.:150.0
Max. :82.58	Max. :5000	Max. :107.0	Max. :91.76	Max. :500.0
NA's :139	NA's :15	NA's :1	NA's :14	NA

Below we see the prevalence of missing data in the dataset. In particular, there is a significant amount of missing data corresponding to the temperature measurement in El Monte, with lesser values of missing data in other variables. Therefore when conducting analysis, the missingness across all covariates will be considered, with special attention on the Temp_ElMonte covariate which has a significant portion of missing data (38%) relative to other measures.



2.2 Analysis Plan

In order to impute the missing data, three major classes of techniques will be used - single imputation, inverse probability weighting (IPW) and multiple imputation. Note that each of these methods rely on the assumption that the missing data is either missing completely at random (MCAR), or missing (conditionally) at random (MAR). Therefore in order to use any of these techniques, a diagnosis of the underlying missing data mechanism must first be established. Though classifying data as MCAR is relatively straightforward through the use of Little's Test (which will be used as the initial step in the analysis), it becomes more complicated when the result indicates the data is not MCAR. If this occurs, the missing data must either be classified as MAR or MNAR, which is harder to verify. This cannot be accomplished from the observed data itself, and therefore typically relies on subject-matter expertise or sampling the missing data (if possible) (Gelman & Hill, 2017).

In the context of this data set, neither of these scenarios are feasible given the data collection method and year. Therefore in the 'worst' case, this report assumes the data to be MAR so imputation can be applied, but first verifies whether or not MCAR is suitable simplification to this assumption. Assuming the data is not MNAR is relatively reasonable in the context of the data set; the primary source of non-random missingness is typically in studies in which information is sensitive and self-reported, neither of which are the case here. Barring extreme chance events (such as natural disasters), there would be no reason to assume non-random missingness in the ozone dataset.

Specifically, the techniques used will be multiple imputation (with m set to be 5, 20 and 40), single imputation (in which the missing observations are replaced with the sample mean for that covariate), inverse probability weighting, and lastly a complete data analysis (in which the missing observations are ignored entirely). Each of the resultant datasets will then be used in a multiple regression model of the ozone variable against the independent covariates, which is a typical model that many non-specialists would often create. In addition to the creation of the models, the efficiency of the regression coefficients will be compared between methods, the resultant model after variable selection, a comparison of the goodness of fit of each model, as well as the differences between point estimations and significance between models.

3 Analysis

3.1 Missing Data Mechanism

Little's Test is initially used to test the validity of the MCAR assumption within the data set. We note that H_0 : missing data is missing completely at random is tested against H_A : missing data is missing at random using a χ^2 distribution. In the context of the Ozone data, we note that we get an extraordinarily small p-value when conducting this hypothesis test, which means we reject the null hypothesis. Therefore we reject the null hypothesis, and conclude (at minimum), the missing data must be missing at random, when examining the entire dataset as a whole.

However, we note that the missing data in all covariates (excluding the Temp_ElMonte covariate) is relatively sparse. Since the data was collected each day over a one year period, it seems inevitable we may eventually run into some missingness as a result of a daily measurement being forgotten, data entry error, a particular measurement device being broken, etc. Therefore for these covariates with sparse missing data, the missing completely at random assumption may be reasonable individually.

Considering there is 38% missingness in the Temp_ElMonte covariate, we may be more concerned with this particular covariate and it's underlying missing mechanism. In order to test this, we fit a logistic regression model in which the outcome is R, which is a binary variable equal to 1 if the observation in Temp_ElMonte is missing, and 0 otherwise.

Table 3: Summary of missing data logistic regression model with missingness in Temp_ElMonte as the response of interest

term	estimate	std.error	statistic	p.value
(Intercept)	-15.269	14.645	-1.043	0.297
Van_Pres	0.003	0.003	1.036	0.300
Wind_Sp	-0.042	0.063	-0.669	0.504
Humid	-0.010	0.010	-0.966	0.334
Temp_Sandburg	-0.011	0.025	-0.430	0.667
Ozone	0.031	0.028	1.123	0.262
Inv_Height	0.000	0.000	-0.347	0.729
Pressure_Grad	0.014	0.007	2.078	0.038
Inv_Temp	-0.010	0.036	-0.276	0.782
Visibility	0.003	0.002	1.506	0.132
SeasonSpring	-0.301	0.444	-0.678	0.498
SeasonSummer	-0.313	0.448	-0.700	0.484
SeasonWinter	-0.116	0.393	-0.295	0.768

From the p-values we can see only one significant covariate, Pressure_Grad, which has a p-value below 0.05. Using a 5% significance level, this appears to be the only covariate significantly related to missingness in the Temp_ElMonte covariate. However, in the context of the data collection itself, we may disregard this relationship. This is because the Pressure_Grad and Temp_ElMonte measurements were taken at entirely different locations within California during the study, and so a relationship between missingness in one observation would likely not depend on another measurement. This significance may just be coincidental given the particular values within the data set.

Overall, our initial assumption that the data was missing at random in the worst case appears to be satisfied in the context of the dataset of interest. Based on both statistical tests and real-world reasoning, it seems appropriate to apply methods relying on MAR as a minimal assumption.

3.2 Single Imputation & Complete Data Analysis

Single imputation is a special case of multiple imputation in which a single imputed estimate is used to replace the missing observations (m =1). As a result of imputing a single value for the missing observations, the resultant estimate is treated as observed with no uncertainty. This is an issue, as in practice, there is often considerable uncertainty associated with the missing data itself, and therefore the standard errors are underestimated in the case of single imputation. The advantage of single imputation relative to complete data analysis is that our overall sample size is enlarged, meaning there is less bias and precision associated with our measurements compared to the list-wise deletion case (Gelman & Hill, 2017).

Many different methods can be employed to impute the missing data. In the simplest case, the analyst can replace the missing observations with the mean of that covariate, though in practice this will often significantly alter the distribution of the covariate and often reduces the correlation between variables. This also relies on the notion that the missing data is similar to the observed data, which depending on the missing data mechanism or context of the problem, may not necessarily be appropriate. A more complex approach relies on modelling the missing covariate in terms of the observed data. In the creation of these models, which may be parametric, semi-parametric, or non-parametric, typically all observed covariates are included in order to improve the predictive accuracy of the final model fit.

The single imputation strategy used in this report will rely on replacing missing observations with their mean. Though this is not theoretically rigorous and has known issues, it remains a popular strategy for non-technical data analysts to approach missing data. This is likely due to the perceived simplicity of both the implementation and logical basis. By using this technique, this report mirrors common pitfalls that may occur as a result of using this method, specifically by comparison of the results from a complete data analysis against single imputation.

In complete data analysis, the estimates are only unbiased if the data is MCAR, which based on the initial investigation, may not be entirely unreasonable. However this is a very strong assumption, and we have no guarantees our observed sample is indeed representative. Therefore this is not intended to be a statistically rigorous method, but rather the output is used as a comparison baseline against other methods to illustrate the effect of different methods at dealing with missing data.

3.2.1 Initial Model Fitting

Table 4: Multiple regression output using single imputation (denoted SI, with missing values replaced by the sample mean) and using a complete data set (denoted CD)

	SI Beta	CD Beta	SI SE	CD SE	SI p-val	CD p-val
Van_Pres	0.000	-0.007	0.005	0.007	0.976	0.290
Wind_Sp	-0.021	0.037	0.125	0.162	0.867	0.821
Humid	0.094	0.118	0.019	0.023	0.000	0.000
Temp_Sandburg	0.207	0.068	0.044	0.066	0.000	0.309
Temp_ElMonte	0.067	0.600	0.041	0.119	0.099	0.000
Inv_Height	0.000	0.000	0.000	0.000	0.414	0.209
Pressure_Grad	-0.028	-0.024	0.012	0.016	0.020	0.132
Inv_Temp	0.117	-0.125	0.063	0.111	0.064	0.258
Visibility	-0.007	-0.003	0.004	0.005	0.064	0.467
SeasonSpring	5.176	5.910	0.833	1.034	0.000	0.000
SeasonSummer	4.203	2.854	0.855	1.091	0.000	0.010
SeasonWinter	0.675	1.246	0.778	0.987	0.386	0.208

From comparing the model in which all missing observations were replaced with the sample mean against the model in which missing observations were ignored, we see a great deal of variability in the results. As a result of the increased sample size used to fit the regression model, we see that relying on the sample mean to fill the missing values has, on average, improved the efficiency of the standard errors in the model. However, we note that the estimates for the coefficients themselves are quite different in some cases - for example the coefficients corresponding to seasons (with fall as the reference category) differ between the two models. Lastly, we note that the significance based on the Wald Test would vary for certain coefficients depending on which analysis strategy we have selected; in particular, the Temp_Sandburg, Visibility and Inv_Temp main effects have large discrepancies in the p-values between the two models.

3.2.2 Model Selection

Table 5: Final model after variable selection using complete data

term	estimate	std.error	statistic	p.value
(Intercept)	-28.96656	2.22996	-12.98971	0.00000
Humid	0.11790	0.01576	7.48278	0.00000
Temp_ElMonte	0.54869	0.03151	17.41328	0.00000
SeasonSpring	6.15216	0.89914	6.84225	0.00000
SeasonSummer	2.40271	0.87131	2.75760	0.00637
SeasonWinter	1.53329	0.91426	1.67708	0.09511

Table 6: Final model after variable selection using single imputation with sample means

term	estimate	std.error	statistic	p.value
(Intercept)	-20.33764	2.33613	-8.70572	0.00000
Humid	0.09710	0.01792	5.41828	0.00000
Temp_Sandburg	0.19264	0.03912	4.92440	0.00000
Temp_ElMonte	0.06232	0.03924	1.58804	0.11317
Pressure_Grad	-0.02810	0.01178	-2.38593	0.01756
Inv_Temp	0.15445	0.03898	3.96242	0.00009
Visibility	-0.00674	0.00344	-1.95766	0.05105
SeasonSpring	5.41340	0.76019	7.12108	0.00000
SeasonSummer	4.27910	0.84402	5.06990	0.00000
SeasonWinter	0.80758	0.75445	1.07042	0.28515

After conducting model selection using the stepAIC function in R, we can see greater volatility in the results of a complete data analysis relative to a model relying on single imputation. In a model reliant on the complete data set only, we see that stepwise model selection results in relatively few parameters in the final model relative to the single imputation strategy. This is as a consequence of the sample size difference between the two methods - in a complete data analysis, the sample size is 203 observations while the single imputation dataset size is 366 observations. The result is smaller standard errors due to this sample size, meaning the significance of covariates is often maintained, and therefore terms are less likely to be dropped from the model.

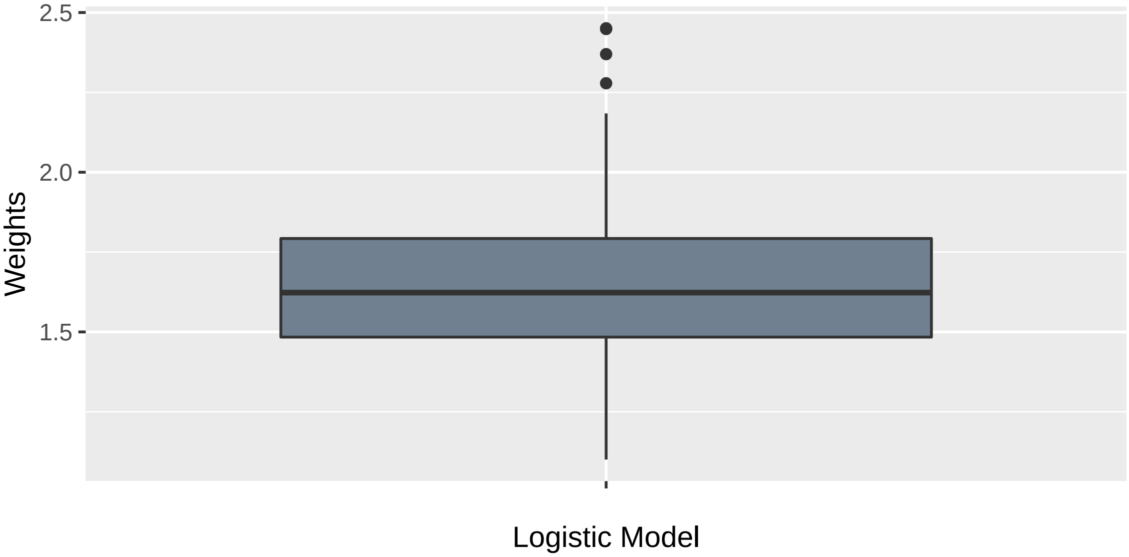
Of the variables maintained in each of the final models, we note that the estimates for the corresponding $\hat{\beta}$ terms are relatively similar between the two models. The magnitude of the seasonal effects are the most different between the two models, though the order of their relative effects is maintained between each model. The adjusted R^2 value is somewhat different between the two models, taking on a value of 0.75 in the complete data analysis and 0.70 in the single imputation model. After conducting model selection, we see that the mechanism in which we approach missing data is important, as the fit, covariates and interpretation of our final model can change considerably.

3.3 IPW

The goal of IPW is to weight our complete observations by the probability of being completely observed (that is, no missingness in any observed covariate). These weights are defined as $w = \frac{1}{\hat{g}_i(y,a,x)}$; $\hat{g}_i(y,a,x)$ denotes the probability of being a complete observation using the observed data, estimated via logistic regression. Similar to the other methods in this paper, IPW relies on a set of assumptions in order to produced unbiased estimators. The first of these assumptions is the missing data is missing at random (though if the data is MCAR, our estimates are still unbiased), and the second is that the missing mechanism is correctly modelled. This second assumption is impossible to verify from observed data alone in practice, and so great caution must be used when defining the weights (Seaman & White, 2015).

In order to estimate the weights, a logistic regression model will be used containing the main effects of each of the covariates within the data set. Other methods, such as generalized boosting models or random forest, could also be used to estimate these weights. However, given the low dimensionality of X in the ozone data, these methods may not provide any significant advantages over logistic regression. Alternate link functions, such as probit or complementary log-log will also be considered.

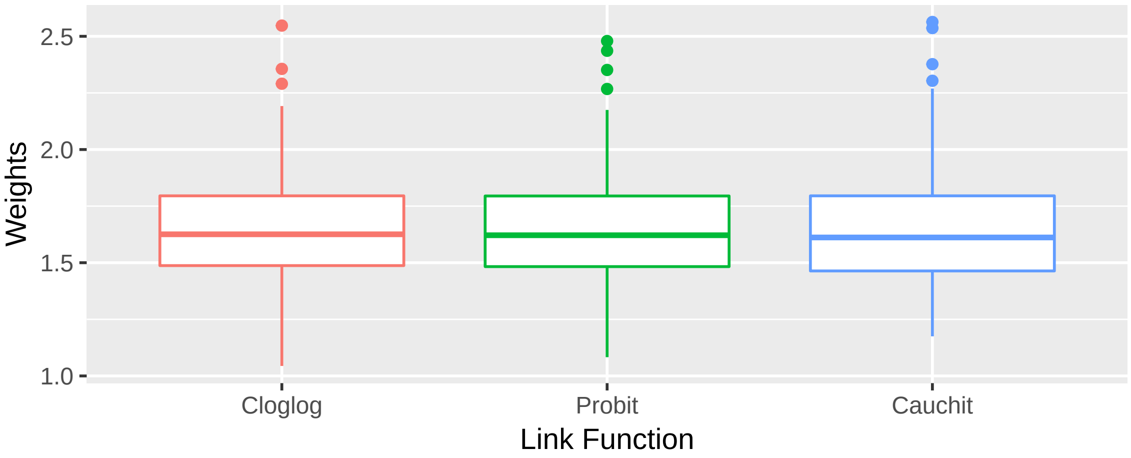
Due to the small proportion of missing data (<5%) in all covariates with the exception of Temp_ElMonte, we note that it is not unreasonable to assume these values are MCAR as previously discussed. Normally to define the weights in IPW, we rely on all covariates with a complete set of observations. In order to make the missing at random assumption more reasonable, it is more sensible to include as many covariates as possible in the model. Therefore we use a simple imputation - the sample mean - to fill in the sparse missing values among the other covariates. The result is that our weights are calculated with all covariates except Temp_ElMonte, which corresponds more closely with the MAR assumption. This also helps increase the overall final sample size.



From the plot above, we can see that the weights are relatively reasonable in the context of the Ozone data set. No weights are so extreme that shrinking is required, and so it appears that logistic regression using a logit link has proven to be effective at weighting the missed observations. If the weights were extreme, alternate strategies such as including interaction terms, conducting variable selection, using different link functions, or using generalized boosting models may also be appropriate. In a more thorough analysis, qualitative research into this data set or more formal modelling could be applied to potentially improve these weights.

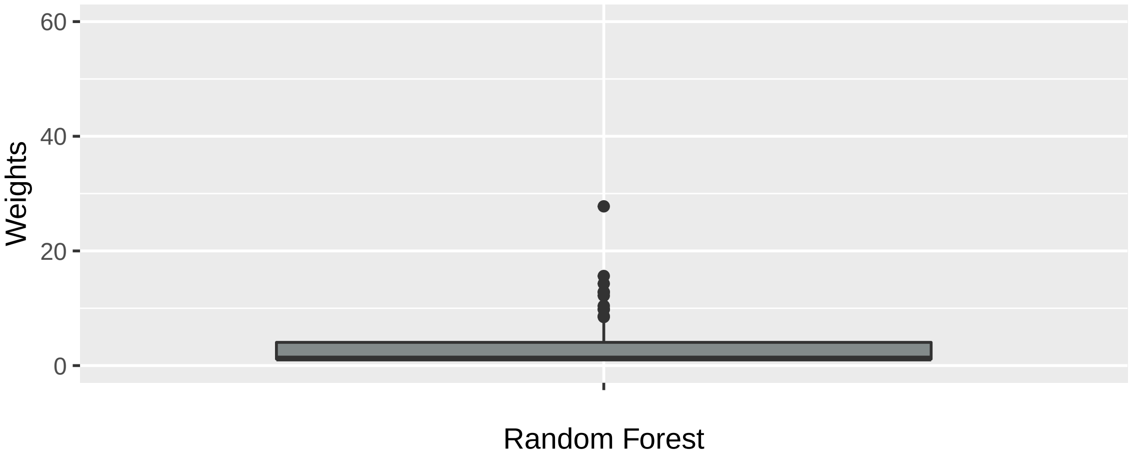
3.3.1 Alternate Strategies

Alternate strategies were also used to compute these weights. Firstly, a probit link and complementary log-log link were used to fit the logistic model, however the results of these estimates were nearly identical (seen in the figure below).



Above we can see the weights are almost identical, regardless of the choice of link function selected in a logistic main effects model. Therefore when using logistic regression, the logit link function is selected for simplicity and interpretability relative to the other options.

A random forest model containing the main effects was also used to compute the weights, as the non-parametric nature of random forest offers more flexibility than a parametric approach. Using a random forest with all main effects, similar to the logistic model, the following weights were computed:



From the plot above we can see the tendency for random forest to produce large weights in the context of this data set. Even after tuning parameters (omitted from the body of this report for brevity), the result was still a tendency of very large weights. This weights would need to be shrunk to avoid certain observations being more heavily weighted than others in the final data analysis, which can lead to issues if a significant portion of weights are above a given threshold (typically 10 or 20).

Therefore from this analysis, it appears a logistic model of the main effects with the logit link is perfectly suitable for this data set, given the relative simplicity of this model and the reasonable estimated weights. Due to the complexity of random forest, and the extreme nature of the weights, this model was not used to define weights in the fitted multiple regression model in subsequent parts. Moving forward, the weights referenced will be the weights computed from the logistic model.

Table 7: Multiple regression model summary after IPW

	Beta Estimations	SE Estimates	p-value
Van_Pres	0.000	0.007	0.977

	Beta Estimations	SE Estimates	p-value
Wind_Sp	0.222	0.144	0.125
Humid	0.124	0.023	0.000
Temp_Sandburg	0.081	0.064	0.207
Temp_ElMonte	0.341	0.113	0.003
Inv_Height	0.000	0.000	0.830
Pressure_Grad	-0.026	0.016	0.112
Inv_Temp	0.059	0.102	0.565
Visibility	-0.003	0.004	0.481
SeasonSpring	6.388	1.067	0.000
SeasonSummer	3.467	1.161	0.003
SeasonWinter	1.963	0.942	0.038

Note that the above model summary comes from the survey package in R; this is necessary in order to correctly account for the variability in the weight estimations, and ensure the associated standard errors for the regression coefficients are correct.

From the results, we can see that IPW has produced similar results to the complete data analysis conducted previously, likely due to the similar sample sizes in each method and the fact that the average weight was not considerably different from 1. Examining the estimates of $\hat{\beta}$, we see these are relatively similar to the previous results in terms of magnitude, with some exception. The coefficients associated with the seasons in particular are the most different relative to before. The standard error associated with the Temp_ElMonte covariate (which contains the majority of the missing data) is nearly identical between IPW and the complete data analysis (0.113 VS 0.119) which may indicate that weighting is not significantly influencing the efficiency for this covariate. Lastly, we observe considerable difference in the associated p-values from the Wald Test between IPW and the complete data analysis, with many of these terms having different significance conclusions relative to before (using $\alpha = 0.05$).

Table 8: IPW model after variable selection

term	estimate	std.error	statistic	p.value
(Intercept)	-27.158	2.511	-10.818	0.000
Humid	0.113	0.015	7.420	0.000
Temp_ElMonte	0.421	0.069	6.110	0.000
Inv_Temp	0.093	0.053	1.768	0.079
SeasonSpring	6.058	0.938	6.460	0.000
SeasonSummer	2.811	0.996	2.823	0.005
SeasonWinter	1.625	0.926	1.755	0.081

After conducting bi-directional stepwise model selection, the final IPW model is similar to the complete data analysis conducted before. Both models share the same main effect covariates, though Inv_Temp is included in the model above, but excluded in the complete data analysis version. The magnitude of each of the coefficients is similar to before, along with the corresponding standard errors for each estimate. Overall from this analysis we can see that introducing the weights to account for missingness did not significantly alter the results relative to the case in which missingness was ignored entirely.

3.4 Multiple Imputation

Multiple imputation offers an improvement over single imputation by creating multiple imputed values for each observations, which offers a better reflection of the underlying variability associated with each imputed value. The number of imputed datasets, represented as m, corresponds to the number of plausible estimates created for each missing value. These estimates are then used to conduct a complete data analysis, and then the results are pooled together through an average calculation to create reasonable point estimates (Grace-Martin, 2018). The advantage of this approach is that Rubin's Rule is used to compute the variability in regression coefficients, which ensures the uncertainty in the imputed values themselves is adequately accounted for in the final output.

In the analysis section, this report will use 3 different values of m in multiple imputation - $m = 5, 20$ and 40. These values were selected in line with previous research, discussed in the earlier sections of this report. The expectation is that larger values of m lead to reduced uncertainty with respect to the imputed values, which should result in a decreased standard error for the resultant regression coefficients in the models. This has the potential to change significance and inferences conducted on the parameters, as well as the overall model fit.

Table 9: Multiple regression output using three values of m in multiple imputation. Note that M1 corresponds to model 1 in which m=5 was used, M2 (model 2) uses m=20, and M3 (model 3) uses m=40

	M1 Coeff.	M2 Coeff.	M3 Coeff.	M1 SE	M2 SE	M3 SE	M1 p-val	M2 p-val	M3 p-val
Van_Pres	-0.01	-0.01	-0.01	0.01	0.01	0.01	0.140	0.084	0.077
Wind_Sp	-0.06	-0.10	-0.08	0.13	0.13	0.13	0.658	0.453	0.532
Humid	0.10	0.10	0.10	0.02	0.02	0.02	0.000	0.000	0.000
Temp_Sandburg	0.15	0.14	0.13	0.06	0.05	0.05	0.009	0.007	0.014
Temp_ElMonte	0.49	0.47	0.50	0.14	0.10	0.11	0.001	0.000	0.000
Inv_Height	0.00	0.00	0.00	0.00	0.00	0.00	0.005	0.005	0.004
Pressure_Grad	-0.04	-0.04	-0.04	0.01	0.01	0.01	0.004	0.003	0.004
Inv_Temp	-0.18	-0.16	-0.17	0.10	0.09	0.09	0.071	0.076	0.067
Visibility	0.00	0.00	0.00	0.00	0.00	0.00	0.268	0.222	0.217
Season_Spring	5.30	5.19	5.22	0.86	0.82	0.83	0.000	0.000	0.000
Season_Summer	4.10	4.07	4.06	0.88	0.85	0.85	0.000	0.000	0.000
Season_Winter	0.87	0.73	0.77	0.80	0.76	0.76	0.279	0.335	0.313

From the table above we see relative homogeneity in the results using the values of $m = 5, 20$ and 40. As expected, we see a slight improvement in the efficiency of the estimates as m increases, since model 3 has resulted in the smallest average standard error between the 3 models. When using the 5% significance level in the Wald Test, we see that we would draw the same conclusions from each model with regard to the significance of each of these main effects. Despite the Temp_ElMonte covariate containing a relatively large proportion of missing values (38%) in the original data set, the estimates are relatively similar across each of the three models. Based on the original data set, these results seem to correspond with expectation. The original data set had relatively small amounts of missing data in all covariates with the exception of Temp_ElMonte, and therefore we would not expect the imputed values to have a significant impact on the overall covariate estimation.

As a result of increasing the number of imputed data sets, we see an improvement in the efficiency of our estimations as a result of using a variety of plausible values for the missing observations. However, the gain in efficiency between 20 and 40 iterations is less than the difference between 5 and 20. This suggests a potential drop-off in efficiency gains after a certain threshold. Using either m as 20 or 40, multiple imputation produces the lowest average standard error across the covariates of all the strategies studied in this report. Therefore based on the relative ease at which this method was applied, the small size of the overall data set, and the robustness of multiple imputation, this appears to be the most suitable mechanism to approach missingness within the ozone data set.

3.5 Comparing Methods

Table 10: Goodness of fit of multiple regression models (with all main effects included) for each missing data strategy

	Rˆ2	AIC
CD	0.76	588.95
SI	0.70	1095.20
MI (m=5)	0.73	1060.10
MI (m=20)	0.73	1061.27
MI (m=40)	0.73	1060.88
IPW	0.76	7583.86

Here we can see the impact of using different approaches to missing data on the overall final fit of each of the multiple regression models. In the case in which missing data is ignored entirely and a complete data analysis is conducted, we achieve the highest R^2 value. This notes the importance of careful consideration of missing data; for a non-technical analyst with limited background in missing data, they may interpret these results as implying ignoring missing data is a suitable analysis strategy given the model goodness of fit. However we know that coefficients used in this model are biased, as the MCAR assumption is likely violated. We also note that different values of m in multiple imputation appear to have a limited effect on the overall R^2 value, though this may change if variable selection was conducted on these models in question. In the context of multiple imputation, the R^2 value was calculated based on the sample median across all m fitted models in each case.

We note that IPW produced the smallest R^2 value, though the AIC is considerably lower than all models aside from the complete data analysis. These results are similar to the single imputation scenario, where we see the highest AIC value of all models examined. Therefore from these results we can see the impact of different approaches to imputing the missing observations on the resultant main effects multiple regression models, though these results do not highlight the severity of potential bias in the estimates of $\hat{\beta}$ in each model.

4 Conclusion

In conclusion, we can see the stark differences in results between each of the missing data strategies. Missing data is a complex issue which requires rigorous investigation into the underlying missing data mechanism, which can be hard to verify in some contexts without previous knowledge of the study design or subject-matter expertise related to the data set. For the Ozone data, in which data was missing across multiple covariates, IPW, single imputation, multiple imputation and a complete data analysis were all explored in order to gain an appreciation for the resultant effects each of these strategies can have on multiple regression.

From these results alone, a non-technical expert may naively conclude that a complete data analysis is the optimal strategy, as it produced relatively small standard errors in the regression coefficients, and had the best model fit (measured through the R^2 and AIC values) of all models. However, if missing data is highly prevalent in the data set (typically >5% is used as a threshold), it must be considered to avoid producing biased estimates. Multiple imputation in particular is a popular strategy to deal with this, as multiple plausible values for the missing observation are considered, and the final model correctly accounts for the variability in these estimates in the final output. In the context of this report, multiple imputation produced reasonable results and appears to be the most suitable missing data approach analyzed. However, this gain in efficiency was not considerable, likely due to the relative low proportion of missingness in the overall data set (6%) and the small relative size of the data set (366 observations).

4.1 Future Research

Future research can be conducted in the following areas:

1) **Model Selection**

In this report, the approach of model selection used was the stepAIC() function, which relies on stepwise variable selection in order to select a final model. This approach was selected intentionally to mirror the typical strategies that may be used by a non-technical data analyst. Despite having known issues, stepwise model selection is pervasive in many fields, likely due to its intuitive nature and simplicity. One must be cautious when interpreting the results of stepwise model selection, as the R^2 values are often inflated, the standard errors can be underestimated, and the p-values used in the Wald Test rely on multiple comparison (Flom, 2018). In a more detailed analysis, other model selection strategies could be investigated to determine the resultant effects.

In each of the regression models, higher order polynomial terms or interaction terms may also be considered if the focus was on finding the optimal fitted model for the data. In the context of the research question, which was to investigate the impact of missing data techniques on a main effects model, these were not necessary. However to be fully comprehensive, these ideas may also be considered in future research and the results could be compared as well.

2) **Logistic Modelling in IPW**

To estimate the weights in inverse probability weighting, the weights were estimated using logistic regression to compute the probability of having a complete observation. In order for the estimates of IPW to remain unbiased, we rely on the assumption that this model is correctly specified. Logistic regression may not be the ideal choice for this estimation, since there is extreme behavior in the tails, and this is a fully parametric method. The former can lead to extreme weights (though this was not an issue in this study), and the latter is a restrictive assumption which is hard to analytically confirm. Other considerations, such as polynomial terms in the main effects, or interactions between covariates may also be considered.

The weights were also briefly estimated using different link functions (probit, cauchit and complementary log log), though the final weights were virtually identical to the logit model. Random Forest was also used to fit the weights, since this is currently a popular technique. The advantage with this approach is that it is a tree-based method, meaning it is non-parametric and therefore relaxes the assumptions of logistic regression. However this is significantly more complex as tuning is required, and different tuning parameters can significantly alter the final results. Even after making attempts to tune the final model, the estimates weights were still on average, higher and more extreme than the simple logistic model case, and so this technique was omitted from the final report.

3) **Multiple Imputation**

In this report, multiple imputation models were constructed using the mice package, which automatically computes variable selection when imputing the missing observations. This underlying mechanism of variable selection, based on relative importance, could be further investigated in future research. Additionally, each of the main effect models used on the imputed data sets were not subjected to any variable selection. This is because variable selection involving multiple imputation is non-trivial, as all of the final models have to be pooled in some way. To my knowledge, no package in R currently exists to efficiently conduct model selection in the context of multiple imputation, so this would require a significant amount of work to do in the future.

5 Appendix A - R Code

```r
#################################################
# Section 1: Data Preprocessing and intialization
#################################################

# Install packages
library(mlbench)
library(gridExtra)
library(grid)
library(ggplot2)
library(lattice)
library(knitr)
library(naniar)
library(mice)
library(BaylorEdPsych)
library(mvnmle)
library(stats)
library(MASS)
library(broom)
library(survey)
library(randomForest)

# Read in data, drop day of the month and week variables
data(Ozone)
drops <- c("V2","V3")
Ozone <- Ozone[ , !(names(Ozone) %in% drops)]

# Rename columns
data.names <- c('Month', 'Ozone', 'Van_Pres', 'Wind_Sp',
                'Humid', 'Temp_Sandburg', 'Temp_ElMonte',
                'Inv_Height', 'Pressure_Grad', 'Inv_Temp', 'Visibility')
colnames(Ozone) <- data.names

# Create new variable for quarters
Ozone$Season <- NA
Ozone$Season[Ozone$Month %in% c(12, 1, 2)] <- "Winter"
Ozone$Season[Ozone$Month %in% c(3, 4, 5)] <- "Spring"
Ozone$Season[Ozone$Month %in% c(6, 7, 8)] <- "Summer"
Ozone$Season[Ozone$Month %in% c(9, 10, 11)] <- "Fall"

# Drop month variable (no longer needed)
Ozone <- Ozone[,-1]

# Create summary table for data set
kable(summary(Ozone[,1:5]), align='c',
      caption = 'Summary of Ozone Data Set (Variables 1-5)')
kable(summary(Ozone[,6:10]), align='c',
      caption = 'Summary of Ozone Data Set (Variables 6-10)')

# Missing data plot
gg_miss_var(Ozone, show_pct = TRUE)
```

```r
# Create summary of missing mechanism using logistic regression
Ozone.missingness <- Ozone
Ozone.missingness$R = ifelse(is.na(Ozone.missingness$Temp_ElMonte),1,0)
missing.mech <- glm(R ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Ozone +
                        Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season,
                    family = 'binomial', data = Ozone.missingness)
kable(tidy(missing.mech), digits=3,
      caption = 'Summary of missing data logistic regression model
      with missingness in Temp_ElMonte as the response of interest')

##################################################
# Section 2: Single Imputation and Complete Data Analysis
##################################################

# Replace missing observations with the sample mean of the variable
Ozone1 <- Ozone
for(i in 1:ncol(Ozone1)){
  Ozone1[is.na(Ozone1[,i]), i] <- mean(Ozone1[,i], na.rm = TRUE)
}

# Fit model using imputed data set
fit.imp <- lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
              Temp_ElMonte + Inv_Height + Pressure_Grad + Inv_Temp +
              Visibility + Season, data=Ozone1)

# Fit model using a complete data analysis
fit.comp <- lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
              Temp_ElMonte + Inv_Height + Pressure_Grad + Inv_Temp +
                Visibility + Season, data=Ozone)

# Create a summary table
comp.sum <- cbind(fit.imp$coefficients, fit.comp$coefficients,
                  summary(fit.imp)$coefficients[,2], summary(fit.comp)$coefficients[,2],
                  summary(fit.imp)$coefficients[,4], summary(fit.comp)$coefficients[,4])
colnames(comp.sum) <- c('SI Beta', 'CD Beta', 'SI SE', 'CD SE', 'SI p-val', 'CD p-val')
kable(comp.sum[-1,], digits = 3,
      caption = 'Multiple regression output using single imputation
      (denoted SI, with missing values replaced by the sample mean)
      and using a complete data set (denoted CD)')

# Conduct bidirectional stepwise selection on each model
# Complete data analysis
step.model1 <- stepAIC(lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
                          Temp_ElMonte + Inv_Height + Pressure_Grad + Inv_Temp + Visibility +
                            Season, data=na.omit(Ozone)), direction='both', trace=FALSE)
# Single imputation
step.model2 <- stepAIC(fit.imp, direction='both', trace=FALSE)

# Print results
kable(tidy(step.model1), digits=5,
      caption = 'Final model after variable selection using complete data')
kable(tidy(step.model2), digits=5,
      caption = 'Final model after variable selection using single imputation with sample means')
```

```r
#################################################
# Section 3: Inverse Probability Weighting
#################################################

# Copy data set
Ozone2 <- Ozone

# Replace missing obs with sample mean
for(i in 1:ncol(Ozone2)){
  Ozone2[is.na(Ozone2[,i]), i] <- mean(Ozone2[,i], na.rm = TRUE)
}

# Overwrite Temp_ElMonte to add back in NA
Ozone2$Temp_ElMonte <- Ozone$Temp_ElMonte
Ozone2$Complete <- as.numeric(complete.cases(Ozone2))
summary(Ozone2)

# Get weights
mechanism <- glm(Complete ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Ozone +
                 Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season,
                 data=Ozone2, family=binomial(link=logit))
unstab.wt = 1/mechanism$fitted

# Create boxplot of weights
unstab.plot <- ggplot(data = data.frame(unstab.wt), aes(x = "", y = unstab.wt, fill=unstab.wt)) +
geom_boxplot(fill='slategrey') + ylab('Weights')

# Create logistic model with each link function

# Probit
mechanism.probit <- glm(Complete ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Ozone +
                        Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season,
                        data=Ozone2, family=binomial(link=probit))

# Complementary log log
mechanism.cloglog <- glm(Complete ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Ozone +
                        Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season,
                        data=Ozone2, family=binomial(link=cloglog))

# Inverse normal CDF
mechanism.cauchy <- glm(Complete ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Ozone +
                        Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season,
                        data=Ozone2, family=binomial(link=cauchit))

# Random Forest
rf.model <- randomForest(as.factor(Complete) ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
                         Ozone + Inv_Height + Pressure_Grad + Inv_Temp + Visibility,
                         data=Ozone2, na.action=na.omit)

# Extract weights for each
cloglog.wts <- mechanism.cloglog$fitted
probit.wts <- mechanism.probit$fitted
cauchy.wts <- mechanism.cauchy$fitted
```

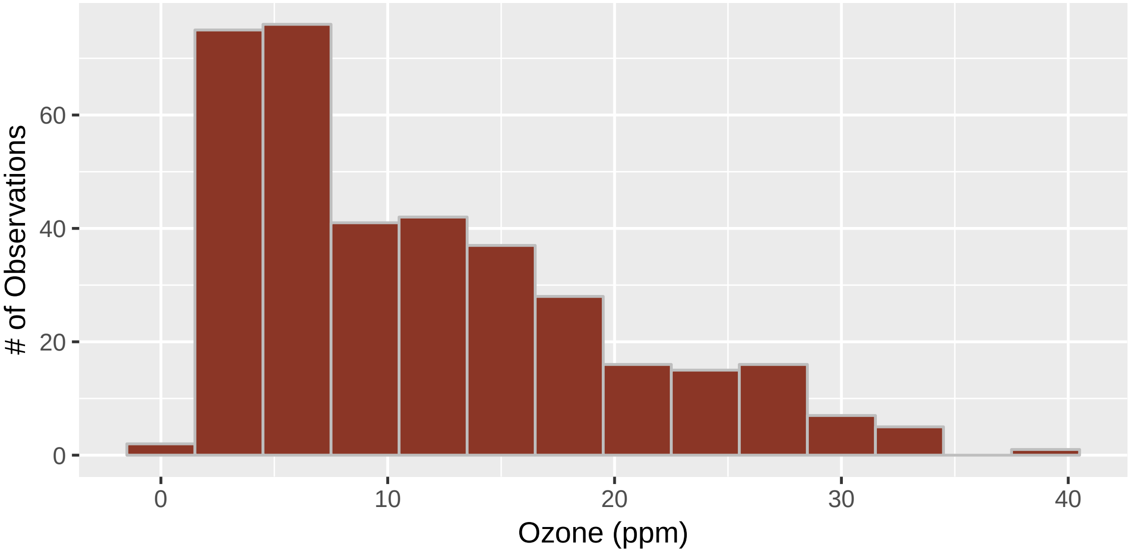
```
rf.wts <- 1/predict(rf.model, Ozone2, type='vote')[,2]
wts <- data.frame(1/cloglog.wts, 1/probit.wts, 1/cauchy.wts)
colnames(wts) <- c('Cloglog', 'Probit', 'Inverse Gaussian')

# Create boxplot of weights by link function
ggplot(stack(wts), aes(x = ind, y = values, color=ind)) +
geom_boxplot() + ylab('Weights') + xlab('Link Function') + theme(legend.position="none")

# Create random forest weight box plot
ggplot(data.frame(rf.wts), aes(x ="", y = rf.wts, fill=values)) +
geom_boxplot(fill='azure4') + ylab('Weights') + ylim(0,60)

# Create IPW model, use survey package to get standard errors
ipw.ps <- svydesign(ids=~1, weights=~unstab.wt, data=Ozone2)
ipw.unstab.model <- svyglm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
                            Temp_ElMonte + Inv_Height + Pressure_Grad +
                            Inv_Temp + Visibility + Season,
                           weights = unstab.wt, design=ipw.ps, data=Ozone2)
summary(ipw.unstab.model)

# Create a summary table for IPW model
ipw.sum <- cbind(ipw.unstab.model$coefficients,
                  summary(ipw.unstab.model)$coefficients[,2],
                  summary(ipw.unstab.model)$coefficients[,4])
colnames(ipw.sum) <- c('Beta Estimations', 'SE Estimates',
                        'p-value')
kable(ipw.sum[-1,], digits=3, caption = 'Multiple regression model summary after IPW')

# Apply stepwise model selection
ipw.unstab.final <- stepAIC(ipw.unstab.model, direction = 'both', trace=FALSE)
kable(tidy(ipw.unstab.final), digits=3, caption = 'IPW, final model')

#################################################
# Section 4: Multiple Imputation
#################################################

# Multiple Imputation - create the 3 imputed data sets
mi1 <- mice(Ozone, m=5, seed=20552745)
mi2 <- mice(Ozone, m=20, seed=20552745)
mi3 <- mice(Ozone, m=40, seed=20552745)

# Create a function which computes the median R squared and AIC for each data set
mult.imp.calc <- function(mi) {
  R.vec <- vector(mode="numeric", length=mi$m)
  AIC.vec <- vector(mode="numeric", length=mi$m)
for (i in 1:mi$m)
  {
lm <- with(complete(mi, action=i), lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg +
                                      Temp_ElMonte + Inv_Height + Pressure_Grad +
                                      Inv_Temp + Visibility + Season))
R.vec[i] <- summary(lm)$r.squared
AIC.vec[i] <- extractAIC(lm)[2]
}
```

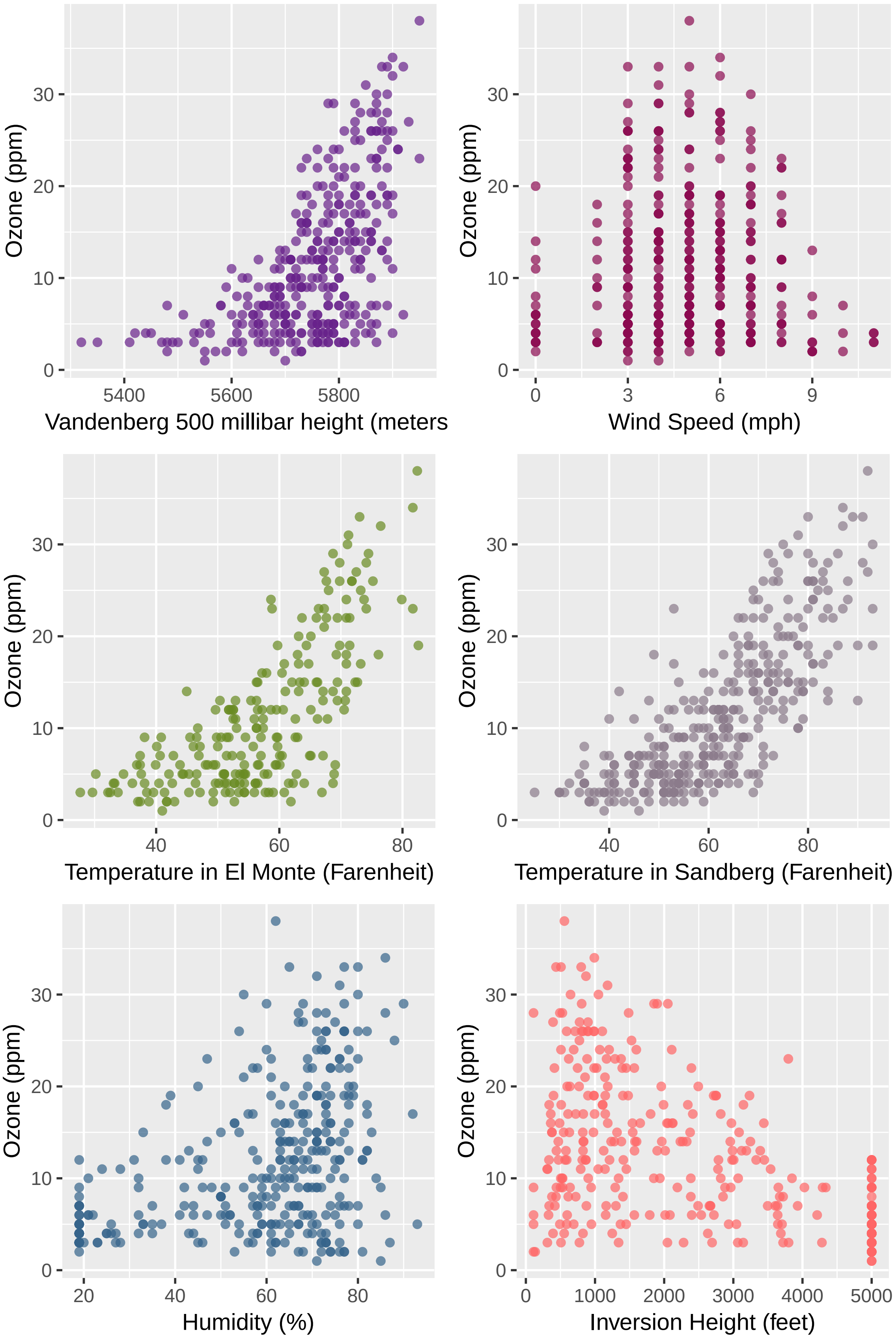
```
  summary.vec <- c(median(R.vec), median(AIC.vec))
  return(summary.vec)
}

# Create the 3 linear models with main effects
fit1 <- with(mi1, lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Temp_ElMonte +
                       Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season))
fit2 <- with(mi2, lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Temp_ElMonte +
                       Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season))
fit3 <- with(mi3, lm(Ozone ~ Van_Pres + Wind_Sp + Humid + Temp_Sandburg + Temp_ElMonte +
                       Inv_Height + Pressure_Grad + Inv_Temp + Visibility + Season))

# Create a summary table
fit1.estimates <- round(summary(pool(fit1))[,1],2)
fit2.estimates <- round(summary(pool(fit2))[,1],2)
fit3.estimates <- round(summary(pool(fit3))[,1],2)

fit1.se <- round(summary(pool(fit1))[,2],2)
fit2.se <- round(summary(pool(fit2))[,2],2)
fit3.se <- round(summary(pool(fit3))[,2],2)

fit1.p <- round(summary(pool(fit1))[,5],3)
fit2.p <- round(summary(pool(fit2))[,5],3)
fit3.p <- round(summary(pool(fit3))[,5],3)

# Create a summary table
mult.imp.sum <- cbind(fit1.estimates, fit2.estimates, fit3.estimates,
                      fit1.se, fit2.se, fit3.se, fit1.p, fit2.p, fit3.p)

# Define table row and column names, and apply them
sum.rows <- c('Intercept', 'Van_Pres', 'Wind_Sp', 'Humid',
              'Temp_Sandburg', 'Temp_ElMonte', 'Inv_Height',
              'Pressure_Grad', 'Inv_Temp', 'Visibility',
              'Season_Spring', 'Season_Summer', 'Season_Winter')
sum.cols <- c('M1 Coeff.', 'M2 Coeff.', 'M3 Coeff.',
              'M1 SE', 'M2 SE', 'M3 SE',
              'M1 p-val', 'M2 p-val', 'M3 p-val')
rownames(mult.imp.sum) <- sum.rows
colnames(mult.imp.sum) <- sum.cols

# Create regression summary table
kable(mult.imp.sum[-1,], align='l',
      caption = 'Multiple regression output using three values of m in multiple imputation.
      Note that M1 corresponds to model 1 in which m=5 was used,
      M2 (model 2) uses m=20, and M3 (model 3) uses m=40')

##################################################
# Section 5: Comparing Models
##################################################

# Create table with AIC and R^2 for each model
fit.rows <- c('CD', 'SI', 'MI (m=5)', 'MI (m=20)', 'MI (m=40)', 'IPW')
fit.cols <- c('R^2', 'AIC')
```

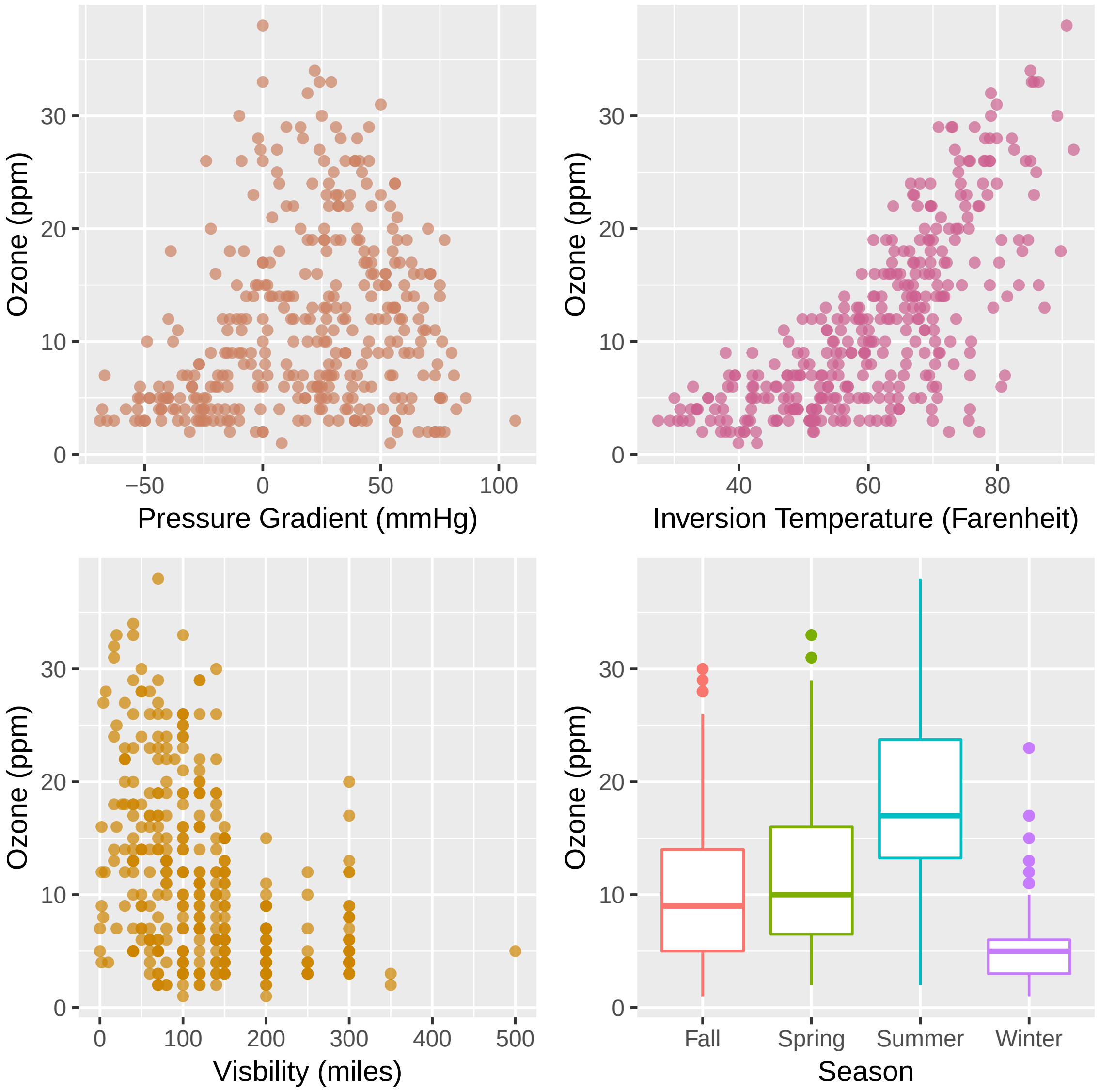
```
R.squares <- c(summary(fit.comp)$r.squared, summary(fit.imp)$r.squared,
               mult.imp.calc(mi1)[1], mult.imp.calc(mi2)[1],
               mult.imp.calc(mi3)[1], summary(ipw.unstab.model)$r.squared)

AIC.vals <- c(extractAIC(fit.comp)[2], extractAIC(fit.imp)[2],
              mult.imp.calc(mi1)[2], mult.imp.calc(mi2)[2],
              mult.imp.calc(mi3)[2], extractAIC(ipw.unstab.model)[2])

fit.table <- cbind(R.squares, AIC.vals)
colnames(fit.table) <- fit.cols
rownames(fit.table) <- fit.rows

kable(fit.table, digits = 2,
      caption = 'Goodness of fit of multiple regression models
      (with all main effects included) for each missing data strategy')

##################################################
# Section 6: Appendix Code
##################################################

# Create plots for each variable against ozone
qplot(Ozone$Ozone,
      geom="histogram",
      xlab = "Ozone (ppm)", fill=I("tomato4"),
      binwidth = 3, ylab = '# of Observations',
      col=I("grey"))

Van_Pres_Plot <- ggplot(Ozone, aes(x=Van_Pres, y=Ozone)) +
  geom_point(color='darkorchid4', alpha = 0.7) +
  xlab('Vandenberg 500 millibar height (meters)') +
  ylab('Ozone (ppm)')

Wind_Speed_Plot <- ggplot(Ozone, aes(x=Wind_Sp, y=Ozone)) +
  geom_point(color='deeppink4', alpha = 0.7) +
  xlab('Wind Speed (mph)') +
  ylab('Ozone (ppm)')

Temp_Sand_Plot <- ggplot(Ozone, aes(x=Temp_Sandburg, y=Ozone)) +
  geom_point(color='thistle4', alpha = 0.7) +
  xlab('Temperature in Sandberg (Farenheit)') +
  ylab('Ozone (ppm)')

Temp_ElMonte_Plot <- ggplot(Ozone, aes(x=Temp_ElMonte, y=Ozone)) +
  geom_point(color='olivedrab4', alpha = 0.7) +
  xlab('Temperature in El Monte (Farenheit)') +
  ylab('Ozone (ppm)')

Humid_Plot <- ggplot(Ozone, aes(x=Humid, y=Ozone)) +
  geom_point(color='steelblue4', alpha = 0.7) +
  xlab('Humidity (%)') +
  ylab('Ozone (ppm)')

InvHeight_Plot <- ggplot(Ozone, aes(x=Inv_Height, y=Ozone)) +
  geom_point(color='#FF6666', alpha = 0.7) +
  xlab('Inversion Height (feet)') +
  ylab('Ozone (ppm)')

Pres_Grad_Plot <- ggplot(Ozone, aes(x=Pressure_Grad, y=Ozone)) +
  geom_point(color='lightsalmon3', alpha = 0.7) +
  xlab('Pressure Gradient (mmHg)') +
  ylab('Ozone (ppm)')

InvTemp_Plot <- ggplot(Ozone, aes(x=Inv_Temp, y=Ozone)) +
  geom_point(color='hotpink3', alpha = 0.7) +
  xlab('Inversion Temperature (Farenheit)') +
  ylab('Ozone (ppm)')

Vis_Plot <- ggplot(Ozone, aes(x=Visibility, y=Ozone)) +
  geom_point(color='orange3', alpha = 0.7) +
  xlab('Visbility (miles)') +
  ylab('Ozone (ppm)')

Season_Plot <- ggplot(Ozone, aes(x=Season, y=Ozone, color=Season)) +
  geom_boxplot() +
  ylab('Ozone (ppm)') + theme(legend.position="none")

# Arrange plots
grid.arrange(Van_Pres_Plot, Wind_Speed_Plot, ncol = 2)
grid.arrange(Temp_ElMonte_Plot, Temp_Sand_Plot, ncol = 2)
grid.arrange(Humid_Plot, InvHeight_Plot, ncol=2)
grid.arrange(Pres_Grad_Plot, InvTemp_Plot, ncol=2)
grid.arrange(Vis_Plot, Season_Plot, ncol=2)
```

6 Appendix B - Data Visualizations

Below are diagnostic plots to characterize the relationship between each of the independent covariate of interest, and their relationship to the dependent variable Ozone.




Ozone (ppm)
0
10
20
30
5400
5600
5800
Vandenberg 500 millibar height (meters
Ozone (ppm)
0
10
20
30
0
3
6
9
Wind Speed (mph)
Ozone (ppm)
0
10
20
30
40
60
80
Temperature in El Monte (Farenheit)
Ozone (ppm)
0
10
20
30
40
60
80
Temperature in Sandberg (Farenheit)
Ozone (ppm)
0
10
20
30
20
40
60
80
Humidity (%)
Ozone (ppm)
0
10
20
30
0
1000
2000
3000
4000
5000
Inversion Height (feet)



Ozone (ppm)
0
10
20
30
−50
0
50
100
Pressure Gradient (mmHg)
Ozone (ppm)
0
10
20
30
40
60
80
Inversion Temperature (Farenheit)
Ozone (ppm)
0
10
20
30
0
100
200
300
400
500
Visbility (miles)
Ozone (ppm)
0
10
20
30
Fall
Spring
Summer
Winter
Season

7 Bibliography

Allison, P. (2012, November 9). Why You Probably Need More Imputations Than You Think. Retrieved April 5, 2019, from https://statisticalhorizons.com/more-imputations

Binding, G., & Willi, T. (2017). A Comparison of Parametric & Non-Parametric Imputation Methods (Unpublished master's thesis). University of Zurich. Retrieved April 11, 2019, from https://ecpr.eu/Filestore/PaperProposal/12136354-7a9f-4094-839d-362748b9f8bd.pdf

Bodner, Todd E. (2008) "What improves with increased missing data imputations?" Structural Equation Modeling: A Multidisciplinary Journal 15: 651-675.

Breiman, L. and Friedman, J. H. (1985), Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association, 80, 580–598.

Donges, N. (2018, February 22). The Random Forest Algorithm. Retrieved April 5, 2019, from https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

Environmental Protection Agency. (2018, September 24). Basic Ozone Layer Science. Retrieved April 5, 2019, from https://www.epa.gov/ozone-layer-protection/basic-ozone-layer-science

Flom, P. (2018, September 23). Stopping stepwise: Why stepwise selection is bad and what you should use instead. Retrieved April 11, 2019, from
https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52

Garson, G. D. (2015). Missing Values Analysis and Data Imputation. Asheboro, NC: Statistical Associates Publishers.

Gelman, A., & Hill, J. (2017). Data analysis using regression and multilevel/hierarchical models.

Grace-Martin, K. (2018, December 15). Missing Data Mechanisms: A Primer. Retrieved April 5, 2019, from https://www.theanalysisfactor.com/causes-of-missing-data/

Graham, John W., Allison E. Olchowski and Tamika D. Gilreath (2007) "How many imputations are really needed? Some practical clarifications of multiple imputation theory." Prevention Science 8: 206–213.

Kang H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64(5), 402–406. doi:10.4097/kjae.2013.64.5.402

Multiple Imputation in a Nutshell. (2018, April 28). Retrieved April 6, 2019, from https://www.theanalysisfactor.com/multiple-imputation-in-a-nutshell/

Nakagawa, S. (2015). Missing data: Mechanisms, methods, and messages. In G. A. Fox, S. Negrete-Yankelevich, & V. J. Sosa (Authors), Ecological statistics contemporary theory and application (pp. 81-105). Oxford: Oxford University Press. doi:10.1093/acprof:oso/9780199672547.001.0001

Schafer, Joseph L. (1999) "Multiple imputation: a primer." Statistical Methods in Medical Research 8: 3-15.

Seaman, S. R., & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research, 22(3), 278–295. https://doi.org/10.1177/0962280210395740

Silverman, B. W. (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. Journal of the Royal Statistical Society, Series B, 47, 1–21.

WeatherStreet. (2010, November 26). Weather Questions & Answers. Retrieved April 5, 2019, from http://weatherstreet.com/weatherquestions/What_is_a_temperature_inversion.htm